# **Natural Language Processing**

Following characteristics of Complex Computing Problem (CCP) are targeted in this semester project of Natural Language Processing

|  |  |
| --- | --- |
|  | Depth of Knowledge |
|  | Depth of Analysis |

**Problem Definition**

This project will involve creating a custom chatbot that answers questions specifically based on information from the Urdu version of the Punjab [Police](https://pkm.punjab.gov.pk/public) Khidmat Markaz (<https://pkm.punjab.gov.pk/public>). You’ll start by crawling the website to gather relevant data and then preprocess this data into a format suitable for the chatbot to retrieve and answer queries accurately.

Your chatbot should be able to:

* Answer user questions strictly based on the data from the crawled content.
* Apologize politely for any queries outside the scope of this data, indicating it’s limited to the specified content.

Using Langchain, you’ll implement conversational memory and retrieval-augmented generation (RAG) to provide accurate, contextually relevant responses while adhering to the scope of the website's information.

**Project Objectives**

1. Develop an understanding of chatbot development.
2. Develop proficiency with Langchain for conversational AI.
3. Create an Urdu dataset from the specified website content(<https://pkm.punjab.gov.pk/public>)
4. Implement data-specific response generation for accurate answers.
5. Build and fine-tune the chatbot to handle in-scope and out-of-scope queries.
6. Evaluate performance through metrics and subjective analysis.
7. Present and defend the approach, implementation, and findings during a viva session.

**Tasks:**

1. **Data Crawling and Preprocessing**
   * Crawl and extract the Urdu content from the specified website (<https://pkm.punjab.gov.pk/public>).
   * Process and clean the data to ensure it’s ready for chatbot integration.
   * Organize the data into a format that enables efficient search and retrieval.
2. **Design Chatbot Workflow**
   * Create a conversation flow that defines how the bot will answer within-scope questions.
   * Develop a fallback flow for handling out-of-scope questions, including a polite apology message.
3. **Develop Initial Prototype**
   * Implement Langchain to build a basic version of the chatbot.
   * Load the processed data into the bot’s memory and enable it to respond to common, in-scope questions.
   * Test the chatbot to ensure that it can correctly identify and apologize for out-of-scope queries.
4. **Testing and Optimization**
   * Conduct extensive testing on a set of user questions to assess accuracy, response time, and scope adherence.
   * Refine the chatbot based on feedback, especially focusing on the accuracy of responses and out-of-scope handling.
5. **Finalize and Present**
   * Complete the chatbot with polished responses and improved conversation flow.
   * Document your process, and prepare a final presentation covering objectives, development journey, challenges, and results.

**Project Deliverables:**

1. **Data Files and Crawling Documentation**: November 19th 2024.
2. **Bot Workflow Design Diagram**: November 26th 2024.
3. **Initial Prototype**: 10th December 2024.
4. **Project Viva**: 24th December 2024.

**Evaluation**

Students will be evaluated on the following Criteria

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Criteria** | **Excellent (90-100%)** | **Good (75-89%)** | **Satisfactory (50-74%)** | **Needs Improvement (<50%)** |
| **Data Collection & Preparation** | Thorough, well-documented data collection and preparation; dataset is highly organized and relevant. | Data collection is complete and organized, with minor gaps in documentation. | Basic data collected with organization; documentation lacks detail. | Incomplete data collection; disorganized or lacks relevance to the project. |
| **Chatbot Design & Workflow** | Workflow is clear, logical, and optimized for user interaction; well-handled in-scope and out-of-scope responses. | Workflow is mostly clear with minor areas for improvement in interaction. | Basic workflow with some gaps in user interaction and response handling. | Unclear or poorly structured workflow; limited or incorrect response handling. |
| **Implementation & Functionality** | Fully functional chatbot; responds accurately to all in-scope queries and gracefully apologizes for out-of-scope ones. | Functional chatbot with minor inaccuracies in response handling. | Partially functional chatbot with occasional incorrect responses or lack of scope control. | Chatbot is largely non-functional or frequently produces incorrect answers. |
| **Language Proficiency & Accuracy** | Responses are accurate, clear, and relevant; language is natural and contextually appropriate. | Responses are mostly accurate, with few errors in language or context. | Responses are somewhat accurate; some language issues or lack of clarity. | Responses are often inaccurate, unclear, or inappropriate in language. |
| **Testing & Evaluation** | Comprehensive testing with detailed feedback and measurable improvements documented. | Testing conducted with feedback; improvements made, though not comprehensive. | Basic testing with minimal feedback; few improvements based on testing. | Insufficient or no testing; limited or no feedback and improvement documented. |
| **Presentation & Documentation** | Presentation is well-organized, insightful, and effectively communicates objectives, methodology, and findings. | Presentation is organized with clear objectives and findings, minor gaps in detail. | Presentation covers main points but lacks depth or organization in some areas. | Presentation is disorganized, incomplete, or lacks clarity in communicating findings. |